

WHAT IS CLAIMED IS:

1. A method of monitoring access to a server system as experienced from multiple user locations on a computer network, comprising:

at a monitoring location, generating first and second request messages that represent requests from users of the server system;

transmitting the first request message over a first communications link from the monitoring location to a first network access location that is remote from the monitoring location for transmission on the network at the first network access location, to thereby simulate access to the server system by a user who is local to the first network access location;

transmitting the second request message over a second communications link from the monitoring location to a second network access location that is remote from both the monitoring location and the first network access location for transmission on the network at the second network access location, to thereby simulate access to the server system by a user who is local to the second network access location; and

at the monitoring location, monitoring responses received from the server system to the first and second request messages;

whereby user access to the server system as experienced by users local to the first and second network access locations is monitored without the need for monitoring components local to the first and second network access locations.

2. The method as in Claim 1, wherein monitoring responses received from the server system comprises:

determining a total response time to the first request message as observed at the monitoring location; and

subtracting from the total response time a delay associated with the first communications link, to thereby determine a response time as experienced at the first network access point.

3. The method as in Claim 2, further comprising estimating the delay based at least on a load applied to the first communications link.

5 4. The method as in Claim 2, wherein the network is the Internet, and the first and second communications links are connection-oriented links that directly connect the monitoring location to the first and second network access locations, respectively.

10 5. The method as in Claim 4, wherein the first and second communications links are Asynchronous Transfer Mode links.

6. The method as in Claim 4, wherein the server system comprises a web server of a publicly accessible web site.

15 7. The method as in Claim 1, wherein the first and second communications links are connection-oriented links that directly interconnect the data center with the first and second network access locations, respectively.

20 8. The method as in Claim 7, wherein the first and second communications links are Asynchronous Transfer Mode links.

9. The method as in Claim 8, wherein the network is the Internet.

25 10. The method as in Claim 1, wherein transmitting the first request message over a first communications link comprises transferring the first request message over the first communications link to a router that is directly connected to the network.

30 11. The method as in Claim 1, wherein transmitting the first request message over a first communications link comprises transferring the first request message over the first communications link to a modem that is locally connected to the network.

12. A system for monitoring performance of a server system as seen from multiple user locations, including user locations that are geographically remote from one another, without the need for performance monitoring components local to the user locations, the system comprising:

5 a plurality of agent computers that are locally interconnected at a central location, the agent computers configured to generate request messages to the server system as simulated users and to monitor responses from the server system to the request messages; and

10 a plurality of communications links, each communications link connecting one or more of the plurality of agent computers to a respective network access point within a respective user location that is geographically remote from the central location;

15 wherein the plurality of agent computer are configured to communicate with the server system via the plurality of communications links to monitor performance of the server system as seen from the multiple user locations.

13. The system as in Claim 12, wherein each communications link is a dedicated, connection-oriented communications link.

20 14. The system as in Claim 12, wherein at least some of the plurality of communications links are Asynchronous Transfer Mode links.

25 15. The system as in Claim 12, wherein at least some of the plurality of communications links produce a round-trip latency that is predicable based on a current load on the communications link.

30 16. The system as in Claim 12, wherein the server system is an Internet server system, and at least some of the communications links are directly peered to the Internet within respective user locations.

17. The system as in Claim 12, wherein the plurality of agent computers are arranged into groups, wherein each group is assigned to a respective communications link and corresponding user location.

18. The system as in Claim 17, wherein the agent computers within a group are configured to use unique IP (Internet Protocol) addresses associated with a corresponding remote Internet point of presence.

19. The system as in Claim 17, wherein at least one of the groups comprises multiple agent computers that are grouped through a hub and a single port of a local switch.

20. The system as in Claim 19, wherein the switch is connected to a router that is directly connected to the plurality of communications links.

21. The system as in Claim 12, wherein the agent computers are configured to measure response times as seen from the user locations by measuring total response times to request messages, and by deducting round-trip latencies associated with the communications links.

22. The system as in Claim 12, further comprising a database that locally stores server performance data generated by the plurality of agent computers.

23. A method for monitoring access to a server system as experienced from a network access location, comprising:

at a first location which is remote from the network access location, generating a user request message that is addressed to the server system;

sending the request message over a communications link to the network access location for transmission over a computer network from the network access location to the server system, to thereby simulate access to the server system by a user who is local to the network access location;

Cont
Sub A3

at the first location, receiving a response to the request message from the server system over the communications link; and
determining a response time to the user request message as experienced at the network access location.

5

24. The method as in Claim 23, wherein determining the response time as experienced at the network access location comprises determining a total response time observed at the first location, and subtracting from the total response time a round-trip delay attributable to the communications link.

10

25. The method as in Claim 24, further comprising estimating the round-trip delay based at least upon a load applied to the communications link.

15

26. The method as in Claim 23, wherein the computer network is the Internet, and the communications link is a connection-oriented link that directly connects the first location to the network access location.

20

27. The method as in Claim 26, wherein the communications link is an Asynchronous Transfer Mode link.

25

28. The method as in Claim 23, wherein the server system comprises a web server of a publicly accessible web site.

29. The method as in Claim 23, further comprising recording the response time, and response times measured for other user locations, within a database that is local to the first location.

30. A system for monitoring access to an Internet server system on a public computer network, comprising:

an agent computer that is configured to access the Internet server system as at least one simulated user while monitoring performance of the Internet server system; and

a dedicated communications link that connects the agent computer to an Internet point of presence that is geographically remote from the agent computer;

wherein the agent computer is configured to use IP (Internet Protocol) addresses associated with the Internet point of presence, such that all forward and reverse message traffic between the agent computer and the Internet server system flows across the dedicated communications link.

31. The system as in Claim 30, wherein the agent computer, or a component locally connected to the agent computer, is configured to measure a total response time to a request message sent to the Internet server system, and to deduct from the total response time a round-trip delay associated with the dedicated communications link, to thereby determine a response time as seen from the Internet point of presence.

32. The system as in Claim 30, wherein the dedicated communications link is a connection-oriented link.

33. The system as in Claim 30, wherein the dedicated communications link is an Asynchronous Transfer Mode link.

34. The system as in Claim 33, wherein the Asynchronous Transfer Mode link is directly peered to the Internet by a router.

35. The system as in Claim 30, wherein the agent computer is one of a plurality of agent computers that are locally coupled to a plurality of dedicated

